Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/IL05/000381

International filing date: 07 April 2005 (07.04.2005)

Document type: Certified copy of priority document

Document details: Country/Office: US
Number: 60/560,050
Filing date: 08 April 2004 (08.04.2004)

Date of receipt at the International Bureau: 22 April 2005 (22.04.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)

PA 1292697

THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE

United States Patent and Trademark Office

March 10, 2005

THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE UNDER 35 USC 111.

APPLICATION NUMBER: ***60/560,050***
FILING DATE: ***April 08, 2004***

By Authority of the
COMMISSIONER OF PATENTS AND TRADEMARKS

H. L. JACKSON
Certifying Officer

PTO/SB/16 (01-04)
Approved for use through 07/31/2006. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PROVISIONAL APPLICATION FOR PATENT COVER SHEET

This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 CFR 1.53(c).

Express Mail Label No.

INVENTOR(S)		
Given Name (first and middle [if any])	Family Name or Surname	Residence (City and either State or Foreign Country)
AMNON	SHASHUA	JERUSALEM, ISRAEL

Additional Inventors are being named on the 1 *separately numbered sheets attached hereto*

TITLE OF THE INVENTION (500 characters max)

PEDESTRIAN DETECTION FOR DRIVING ASSISTANCE SYSTEMS

Direct all correspondence to: **CORRESPONDENCE ADDRESS**

☐ Customer Number:

OR

☑ Firm or Individual Name	MOBILEYE TECHNOLOGIES LIMITED				
Address	IDLIA HOUSE				
Address	3 THEMISTOKLI DERVIS STREET				
City	NICOSIA	State		Zip	CY-1066
Country	CYPRUS	Telephone	+357 22 671-808	Fax	+357-22-675 446

ENCLOSED APPLICATION PARTS ***(check all that apply)***

☑ Specification *Number of Pages* 9 ☐ CD(s), Number _____

☐ Drawing(s) *Number of Sheets* _____ ☐ Other (specify) _____

☐ Application Data Sheet. See 37 CFR 1.76

METHOD OF PAYMENT OF FILING FEES FOR THIS PROVISIONAL APPLICATION FOR PATENT

☑ Applicant claims small entity status. See 37 CFR 1.27.

☑ A check or money order is enclosed to cover the filing fees.

FILING FEE Amount ($): $80-

☐ The Director is herby authorized to charge filing fees or credit any overpayment to Deposit Account Number: _____

☐ Payment by credit card. Form PTO-2038 is attached.

The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.

☑ No.

☐ Yes, the name of the U.S. Government agency and the Government contract number are: _____

Respectfully submitted,

SIGNATURE Gdalyahu Yoram

TYPED or PRINTED NAME YORAM GDALYAHU

TELEPHONE +972-2-541-7333

Date 03/22/04

REGISTRATION NO. _____ *(if appropriate)*

Docket Number: _____

USE ONLY FOR FILING A PROVISIONAL APPLICATION FOR PATENT

This collection of information is required by 37 CFR 1.51. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 8 hours to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Mail Stop Provisional Application, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

PROVISIONAL APPLICATION COVER SHEET
Additional Page

PTO/SB/16 (08-03)
Approved for use through 07/31/2006. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

Docket Number

INVENTOR(S)/APPLICANT(S)		
Given Name (first and middle [if any]	Family or Surname	Residence (City and either State or Foreign Country)
YORAM	GDALYAHU	JERUSALEM, ISRAEL
GABY	HAYUN	JERUSALEM, ISRAEL

Number 1 of 1

WARNING: Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.

.01919 U.S. PTO

PTO/SB/17 (10-03)
Approved for use through 07/31/2006. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

FEE TRANSMITTAL for FY 2004

Effective 10/01/2003. Patent fees are subject to annual revision.

☑ Applicant claims small entity status. See 37 CFR 1.27

TOTAL AMOUNT OF PAYMENT ($) 80

Complete if Known	
Application Number	
Filing Date	
First Named Inventor	
Examiner Name	
Art Unit	
Attorney Docket No.	

METHOD OF PAYMENT *(check all that apply)*

☑ Check ☐ Credit card ☐ Money Order ☐ Other ☐ None

☐ Deposit Account:

Deposit Account Number
Deposit Account Name

The Director is authorized to: *(check all that apply)*

☐ Charge fee(s) indicated below ☐ Credit any overpayments

☐ Charge any additional fee(s) or any underpayment of fee(s)

☐ Charge fee(s) indicated below, **except for the filing fee** to the above-identified deposit account.

FEE CALCULATION

1. BASIC FILING FEE

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description	Fee Paid
1001	770	2001	385	Utility filing fee	
1002	340	2002	170	Design filing fee	
1003	530	2003	265	Plant filing fee	
1004	770	2004	385	Reissue filing fee	
1005	160	2005	80	Provisional filing fee	80

SUBTOTAL (1) ($) 80

2. EXTRA CLAIM FEES FOR UTILITY AND REISSUE

	Extra Claims		Fee from below		Fee Paid
Total Claims		-20** =		X	=
Independent Claims		- 3** =		X	=
Multiple Dependent					=

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description
1202	18	2202	9	Claims in excess of 20
1201	86	2201	43	Independent claims in excess of 3
1203	290	2203	145	Multiple dependent claim, if not paid
1204	86	2204	43	** Reissue independent claims over original patent
1205	18	2205	9	** Reissue claims in excess of 20 and over original patent

SUBTOTAL (2) ($)

***or number previously paid, if greater; For Reissues, see above*

FEE CALCULATION (continued)

3. ADDITIONAL FEES

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description	Fee Paid
1051	130	2051	65	Surcharge - late filing fee or oath	
1052	50	2052	25	Surcharge - late provisional filing fee or cover sheet	
1053	130	1053	130	Non-English specification	
1812	2,520	1812	2,520	For filing a request for *ex parte* reexamination	
1804	920*	1804	920*	Requesting publication of SIR prior to Examiner action	
1805	1,840*	1805	1,840*	Requesting publication of SIR after Examiner action	
1251	110	2251	55	Extension for reply within first month	
1252	420	2252	210	Extension for reply within second month	
1253	950	2253	475	Extension for reply within third month	
1254	1,480	2254	740	Extension for reply within fourth month	
1255	2,010	2255	1,005	Extension for reply within fifth month	
1401	330	2401	165	Notice of Appeal	
1402	330	2402	165	Filing a brief in support of an appeal	
1403	290	2403	145	Request for oral hearing	
1451	1,510	1451	1,510	Petition to institute a public use proceeding	
1452	110	2452	55	Petition to revive - unavoidable	
1453	1,330	2453	665	Petition to revive - unintentional	
1501	1,330	2501	665	Utility issue fee (or reissue)	
1502	480	2502	240	Design issue fee	
1503	640	2503	320	Plant issue fee	
1460	130	1460	130	Petitions to the Commissioner	
1807	50	1807	50	Processing fee under 37 CFR 1.17(q)	
1806	180	1806	180	Submission of Information Disclosure Stmt	
8021	40	8021	40	Recording each patent assignment per property (times number of properties)	
1809	770	2809	385	Filing a submission after final rejection (37 CFR 1.129(a))	
1810	770	2810	385	For each additional invention to be examined (37 CFR 1.129(b))	
1801	770	2801	385	Request for Continued Examination (RCE)	
1802	900	1802	900	Request for expedited examination of a design application	

Other fee (specify) _____

*Reduced by Basic Filing Fee Paid **SUBTOTAL (3)** ($)

SUBMITTED BY

					(Complete *(if applicable)*)
Name (Print/Type)	YORAM GDALYAHU	*Registration No. (Attorney/Agent)*		*Telephone*	+972-2-541-7333
Signature	Gdalyahu Yoram			*Date*	03/22/04

WARNING: Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.

This collection of information is required by 37 CFR 1.17 and 1.27. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Pedestrian Detection for Driving Assistance Systems

Amnon Shashua
Hebrew University of Jerusalem
School of Eng. and CS
Jerusalem, 91904 Israel

Yoram Gdalyahu
MobilEye Ltd.
R.M.P.E. build., Har-Hotzvim
Jerusalem, 91450 Israel

Gaby Hayun
MobilEye Ltd.
R.M.P.E. build., Har-Hotzvim
Jerusalem, 91450 Israel

Abstract

We describe the functional and architectural breakdown of a monocular pedestrian detection system. We describe in detail our approach for single-frame classification based on a novel scheme of breaking down the class variability by repeatedly training a set of relatively simple classifiers on clusters of the training set. Single-frame classification performance results and system level performance figures for daytime conditions are presented with a discussion about the remaining gap to meet a daytime normal weather condition production system.

1 Introduction

This paper describes a monocular visual processing system for pedestrian detection targeting the niche of driving assistance on-board vehicles. The development is geared towards a serial production sensor qualified initially for collision warning and ACC Stop & Go applications, and later for active safety collision mitigation systems. The system runs on a prototype development platform (based on 1GHZ microprocessor PPC7457 G4) at a rate of 10Hz and is being ported onto a specialized system-on-a-chip (EyeQ) with the target frame rate of 20-25HZ.

Generally speaking, a visual processing system needs to function well under a wide range of visibility conditions covering over-cast sky, strong highlights, low visibility due to inclement weather, wide dynamic range of imaging conditions, change of context, day-time and night-time driving. On top of that, the class of pedestrians is particularly challenging for a number of reasons:

- The image space variability of the class is very large as pedestrians appear in various poses, clothing and various articulations of body parts. The articulation of body parts also makes the process of tracking a pedestrian along an image sequence somewhat challenging.
- Pedestrians are found mostly in city traffic conditions where the background texture (from surrounding man-made structures, other vehicles poles and trees) form a highly cluttered environment.
- The background clutter covers both shape (texture) and depth. If in an open roadway a pedestrian would stand out using depth disparity cues (such as by using stereopsis), depth cues are unlikely to be useful for segmenting out pedestrians in city traffic due to the heavy disparity clutter.
- Pedestrians occupy a narrow image strip and from a distance may look similar to many background objects such as trees, poles, parts of parked vehicles, narrow windows and openings, and so forth.
- Laterally moving pedestrians form an important sub-class for which motion measurements form a powerful cue. However, parts of moving vehicles (in slow traffic) also generate inward motion signals and motion-based segmentation from a moving platform is still a difficult problem especially in an evironment rich with other moving structures.

We will present below the functional and architectural breakdown of a pedestrian detection system, and in more details present a novel single-frame detection algorithm. One of the key points which emerges from our analysis

of single-frame detection schemes is that it is unrealistic to expect a reasonable system level performance using single-frame classification only. Only by pooling together many perceptual decisions can the system hope to segment out pedestrians at a sufficiently reliable level. The key therefore lies in the integration of additional cues measured over time (dynamic gait, motion parallax, stability of re-detection measures), situation specific features (such as leg positions at certain poses), and most importantly via building up additional object categories consisting of vehicles (both in motion and stationary) and stationary background structure such as poles, trees, guardrails, lane markings and so forth. Due to space limitations we will focus on the details of the single frame detection and describe only in general terms the principles of the multi-frame decisions and end with a detailed comparative analysis of our classifier and present results and statistics of the system level performance.

2 Functional Breakdown of the System

The appearance of pedestrians in the scene can be divided into a number of categories:

Pedestrians moving laterally: visual motion analysis is a strong cue for detection provided that the host vehicle motion is factored out. In other words, simple image subtraction would not apply since the camera is mounted on a moving platform. Another important cue is the gait pattern both dynamically (change of position of legs over time) [3, 17, 4] and statically (position of legs in a single frame).

Stationary pedestrians in primary host vehicle path ("in-path"): pattern recognition (based on texture/shape) is the primary source. In some cases the static gait position becomes useful. Motion parallax from the ground plane [13, 11] forms a weak cue for sufficiently close pedestrians, however it might be difficult to extract reliably in a city traffic environment due to clutter and low visibility of the roadway due to occlusions.

Pedestrians moving longitudinally: similar to stationary pedestrians as the longitudinal motion is too weak to be reliably picked up by image processing.

Stationary pedestrians out-of-path: stationary pedestrians out of the host vehicle path need to be detected in order to minimize the detection delays in case the pedestrian decides to move inwardly — no external action is expected from the system upon detection of stationary out-of-path pedestrians thus a certain level of false positives is allowed.

The pedestrian system architecture loops through the following modules:

(1) **Generate candidate regions of interest:** a systematic scan of the image for rectangular shaped regions at all positions and all sizes would be computationally unwieldy. An attention mechanism filters out windows based on lack of distinctive texture properties and incompliance with perspective constraints on range and size of the candidate pedestrian. On average, the attention mechanism generates 75 windows (out of the many thousands of candidates which could be generated otherwise) per frame which are fed to the classifier.

(2) **Single frame classification:** this is the heart of the detection process. Details are presented in Section 3.

(3) **Multi-frame Approval Process:** candidates which survive the single frame classification thresholds are likely to correspond to pedestrians . However, due to the high variability of the object class and the high levels of background clutter it is conceivable that coincidental arrangements of image texture may have a high detection score — an ambiguous situation which is likely to be unavoidable. Additional information collected over a number of frames are used in the system for further corroboration. Measures that are collected over multiple frames include (i) dynamic gait pattern based on periodicity, (ii) inward motion analysis scores (coupled with ego-motion [14]), (iii) motion parallax (when available), (iv) consistency measure of the single-frame classifier over time, and (v) tracking quality measures. The approval process is based on a decision-tree type of classifier trained by a training set. The length (number of frames) of the approval process depends on the type and quality of the collected information. For example, a strong inward motion ranks highly in the decision process and induces an immediate approval.

(4) **Range measurement:** candidate regions are fit to pedestrians in such a way that the lower part of the rectangular region is aligned with the feet. A gait recognition process is used both as a discriminant in the single-frame classification and multi-frame approval and for a cue for range measurement. More details on the process of range measurement using the flat roadway assumption can be found in [15].

The four basic steps above are also coupled with supporting functions such as host vehicle ego-motion (of Yaw and Pitch) [14], close range motion segmentation (for extracting strong inward motion regardless of shape classification), robust tracking (which can handle non-rigid motion and occlusions induced by pedestrians crossing each other) — and of primary importance the classification scores of background sub-classes which include licensed vehicles, poles, guard-rails, repetitive texture, lane mark interpretation, bridges and other man-made horizontal structures, and pedestrian walking zone areas. The sub-class scores play an important role in the final decision-tree multi-frame approval process.

We describe next our approach for single-frame classification and present a novel scheme designed to reduce the class variability to smaller pieces by repeatedly training a set of relatively simple classifiers on clusters of the training set.

3 Single Frame Classification Algorithm

The changing pose and articulation of the limbs suggests a classifier based on the integration of local image representations as opposed to a holistic (global) representation. A local image representation breaks down the class variability to local parts each with its own variability which is presumably much smaller than that of the entire shape. Moreover, the representation by components compensate for pose and articulation changes by allowing a flexible geometric relation among the components during classification. The integration of the local representations in the classification stage can be rather degenerate such as the nearest neighbor approach which employs relatively sophisticated local features such as those used by [9], or integration via a cascaded classifier such as the hierarchical SVM approach used by [8].

Both local feature integration approaches are problematic in our application domain. The nearest neighbor approach has been proven very effective for matching against a single exemplar (as opposed to a class of objects), when the number of descriptors is relatively large (in the thousands) and when the local descriptors are localized on richly textured regions [9]. In our application domain, the image regions surrounding a typical pedestrian are often poorly textured and because of the small region size it is difficult to generate a large number of distinct descriptors. The hierarchical SVM runs an SVM [1] classifier separately on each local region thereby mapping each sub-region to a real number (distance to local decision surface) — which can be considered as a local discriminant function. The results of the local discriminants are integrated by running an SVM classifier on the feature vector comprising of the local discriminant results. Due to the relatively small number of local regions (of the order of 10), one would require each sub-region to be highly discriminatory, be localized in order to maximize the discrimination ability and to be subject to a relatively sophisticated component classifier which in the context of SVM translates to a high order feature map (polynomial of Radial Basis Function). The number of support vectors (templates used during the classification stage) for high order feature maps are relatively large (roughly 10% of the training set) thus make the classification stage costly in computing resources. In [8] a quadratic polynomial component classifier is used where the integration of local discriminants is done by a linear classifier. In our domain, mainly due to the small image size of interest regions (candidate regions are warped to a 12×36 window which is fed to the classifer) and the poorly defined sub-regions which make accurate localization of component regions very challenging, such an approach is not strong enough for an effective single-frame classifier (see comparative results in Section 4).

3.1 Multi-training Classification by Components: Our Approach

Our approach to the single-frame classification stage borrows from the idea of the recognition-by-components using a 2-stage classifier algorithm. Namely, we breakdown the region of interest into sub-regions, create a local vector representation per sub-region, feed each of the local feature vectors to a discriminant function and integrate the local discriminant results by a second-stage classifier. The crucial difference from the conventional paradigm is the way we handle the training set. Since the number of local sub-regions are small we generate multiple local discriminants (one per local sub-region) by dividing the training set into mutually exclusive *training clusters* where each cluster represents a training collection from a particular pose, a particular articulation and a

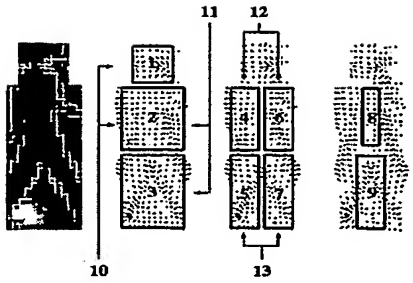

Figure 1: *The configuration of the nine sub-regions is displayed over the gradient image. The distribution of the arrows in each sub region is measured (see the text). In addition, four pair combinations are constructed (regions 10, 11, 12, 13).*

particular illumination condition — all together 9 different subsets of the training dataset. The idea behind the subset division of the training set is to breakdown the overall variability of the class into manageable pieces which can be captured by relatively simple component classifiers. In other words, rather than seeking sophisticated component classifiers which cover the entire variability space (of the subregions) we apply prior knowledge in the form of clustering (manually) the training set. Each component classifier is trained multiple times — once per training cluster — while the multiple discriminant values per subregion and across subregions are combined together via Adaboost [5]. Details of this approach are given below.

The candidate region is divided into a fixed configuration of 9 overlapping sub-regions with positions illustrated in Fig. 1). We next compute for each subregion a local image descriptor designed to be insensitive to local shifts of image structure that may be caused by change of pose and articulation of the pedestrian's arms and limbs. The particular design of the descriptor vector is borrowed from a biologically inspired model implemented by [9] following necessary changes due to the relatively small size of our sub-regions. In particular, image gradient magnitudes and orientations are sampled and the robustness against local shifts is achieved by creating orientation histograms over 2×2 sample regions (i.e., the sub-region is further divided in a 2×2 configuration). Each orientation histogram has 8 orientation bins whose level are weighted by a smoothed version of the gradient magnitudes. Taken together, the local description consists of $2 \times 2 \times 8 = 32$ element feature vector normalized to unit length (the normalization reduces the effects of illumination changes).

The 32-element feature vector (per sub-region) undergoes a linear weighting using Ridge Regression [7]. Briefly, let $\mathbf{w} \in R^{32}$ be the desired weight vector which ideally forms a hyperplane which separates the negative and positive examples of the training cluster (the process below is applied separately to each training cluster). Let $\mathbf{x}_i$ be the input local descriptors corresponding to the i'th training image (at the particular location of sub-region) and let $y_i = \pm 1$ denote the class label. The Ridge Regression procedure seeks $\mathbf{w}$ which minimizes the objective function:

$$\alpha \|\mathbf{w}\|^2 + \sum_i (y_i - \mathbf{w}^\top \mathbf{x}_i)^2,$$

where α is a pre-determined fixed positive constant. The solution of the optimization problem can be described in closed form as follows. Let M be the Gram matrix, i.e., $M_{ij} = \mathbf{x}_i^\top \mathbf{x}_j$, let $A = [\mathbf{x}_1, \mathbf{x}_2, ...]$ hold the input vectors as its columns, and let $\mathbf{y} = (y_1, y_2, ...)$. The weight vector $\mathbf{w}$ is equal to:

$$\mathbf{w} = A(K + \alpha I)^{-1}\mathbf{y}.$$

The discriminant of $\mathbf{x}$ is the inner-product $\mathbf{w}^\top \mathbf{x}$. Given a particular training cluster, the inner-product between the 9 weight vectors (one per sub-region) and their corresponding sub-region local descriptor vector form a feature vector of 9 elements. Four additional elements are added by concatenating selected *pairs* (illustrated in Fig. 1) of sub-regions into local descriptors with 64 elements each which are turned into 4 elements by the Ridge Regression procedure above. Taken together, we have 13 elements per training cluster, thus making a single feature vector of $9 \times 13 = 117$ elements representing the candidate region.

Note that breaking apart the training set into clusters is a crucial ingredient in this procedure, because otherwise the linear discriminant (per sub-region) would be too weak of a classifier to be of practical use. Our findings show that simplifying the variability space (induced by the training clusters) is much more powerful than seeking a stronger local discriminant — most likely because the local image structure is not sufficiently discriminatory for such a wide variability space. Also note that the choice of the local descriptor allows us to bypass the need for *localized* features, i.e., searching for distinguishable parts such as arms, legs, head, and so forth.

The 117 elements are combined with Adaboost using the *entire* training set. Each of the 117 elements can be considered as a "weak lerner" in the sense that it forms a class discrimination. The main idea of AdaBoost is to assign each example of the training set a weight. At the beginning all weights are equal, but in every round the weak learner returns a hypothesis, and the weights of all examples classified wrong by that hypothesis are increased. That way the weak learner is forced to focus on the difficult examples of the training set. The final hypothesis is a combination of the hypotheses of all rounds, namely a weighted majority vote, where hypotheses with lower classification error have higher weight.

This completes the description of the 2-stage classification algorithm. In the next section we compare our approach to holistic SVM and 2-stage SVM and demonstrate a significant improvement in the ROC curve.

4 Experimental Results

The single frame pedestrian classification phase has been a subject of past research (cf. [2, 6, 8, 12]) with published performance figures. In general, the performance of any classification system is subject to a tradeoff between the rate of miss-detections (false negatives) and the rate of false detections (false positives). For example, the performance of detection drops as one imposes more stringent restrictions on the rate of false positives. This tradeoff is captured by the so called ROC curve which plots the error in miss-detection against the false alarm rate.

The images were captured at a 640×480 resolution with a horizontal field of view of 47 degrees. Regions of interest were scaled (up or down) to fill a region of size 12×36 pixels which were fed into the single-frame classifer. The training dataset consisted of 54,282 instances roughly split equally between positive (pedestrians) and negative examples. The negative examples were generated automatically by sampling the windows produced by the system's attention mechanism. It is important to note that the attention mechanism filters out windows based on lack of distinctive texture properties and incompliance with perspective constraints on range and size of the candidate pedestrian. In other words, the negative examples are not random image fragments. On average, the attention mechanism generates 75 windows per frame which are fed to the classifier. The test dataset consisted of 15,244 instances, where both the training and test sets cover a wide variety of daytime conditions including scale (range to camera from 3m to 25m), pose, articulation, illumination, background texture, weather conditions, and a spectrum of visibility conditions (mostly due to inclement weather conditions). The training and test sets were extracted from 50 hours of driving covering city traffic conditions around the world including Japan, Munich, Detroit and Israel.

Fig. 2 shows the ROC curve (the top curve) of the *test* dataset. We can see from the curve that our classifier would achieve, for example, a detection rate of 90% at a false rate of 5.5% (which means 1 false positive for every 18 windows inspected). We chose the tradeoff with 93.5% detection rate — achieved at a false positive rate of 8% which is roughly one false positive for every 12 windows inspected. Fig. 3 shows a sample of false positives (upper row) and false negatives of the single frame classification phase. Note that a window containing a pedestrian at a wrong scale is also considered a false positive since the system cannot reliably track these regions over time and thus the region will be eventually dropped during the multi-frame classification phase.

The reader may notice that these figures are strikingly poorer than previously published results. For example, [12] applies a global SVM on windows of size 64×128 and reports a detection rate of 81.6% at a false positive rate of 1 window per 15,000 windows inspected. The difference in reported results may arise from a number of sources: (i) the test set in [12] consisted of 165 positive examples only, (ii) negative examples were generated by systematically scanning the image — therefore many of the negative examples were "very easy", as opposed to the negative examples generated by our attention mechanism, (iii) different window size of 64×128 suggests that high detailed pictures of pedestrians were used as opposed to the often impoverished images our system must

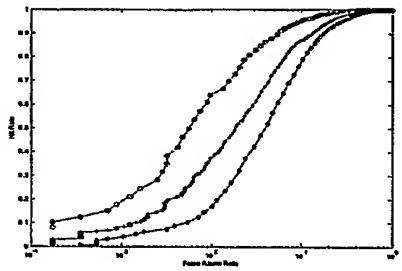

Figure 2: *ROC curves of three classifiers: our classifier is the top curve, the global SVM classifer [12] is the middle curve, and the 2-stage SVM classifier [8] is the bottom curve.*

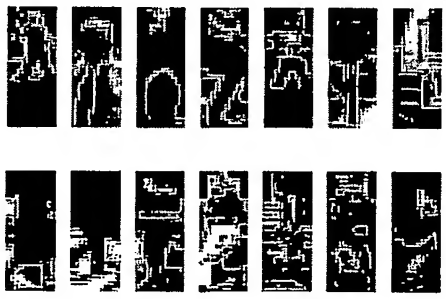

Figure 3: *Some misclassifiaction examples. Upper row: false positive examples. Bottom row: false negative examples.*

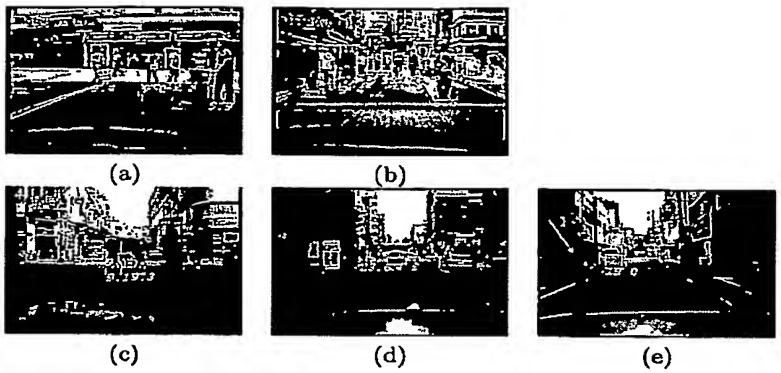

Figure 4: *(a) and (b) illustrate a typical image from the 1 hour test ride in one of the busiest Tokyo districts. The numbers below each bounding box represent range. (c),(d) contain false positive examples: the horse legs in (c) and what may appear as an out-of-path pedestrian at a range of 23m in (d). The region size at a distance of 23m is roughly* 8×24 *pixels. Square like bounding box represents a vehicle in (d) and (e). An example of a miss detection (where the arrow is located) in (e).*

handle, (iv) training and test sets in [12] cover only rear and front poses whereas in our case all poses are covered, and (v) it is unclear what level of variability (i.e., degree of challenging situations) are covered by a particular dataset and thus it becomes difficult to make comparisons. Similar performance figures on the component-based 2-stage SVM were reported in [8] citing an ROC curve with a 90% detection rate with a false positive rate of 1 to 10,000 windows inspected.

Since it is difficult to establish a baseline for comparing the various approaches from the published performance figures alone, we have run the global SVM and a 2-stage SVM using our training and test tests — thus establishing a common baseline for comparative performance evaluation of the single-frame classification phase. The middle ROC curve in Fig. 2 corresponds to a quadratic polynomial classifier trained using SVM (with the quadratic kernel function) using the procedure described in [12]. The lower ROC curve corresponds to a 2-stage SVM (following [8]) where the 9 sub-regions and 4 pairs of sub-regions were classified using a quadratic polynomial SVM and the components combination was done with a linear SVM. The reader may wonder why the component 2-stage SVM performs poorer than the global SVM — contrary to the results of the original authors. The reason lies, in our opinion, with the poor texture definition of the small sub-regions which make each of them not sufficiently discriminatory over the entire variability space. This underscores the importance of our approach to breakdown the variability to smaller pieces thus enabling the small sub-regions to become sufficiently discriminatory. In comparison, the sub-regions in [8] were relatively large (ranging from 28×28 to 69×46) thus allowing the underlying image texture to become more discriminatory. In anycase, both ROC curves are uniformly poorer that the ROC curve of our classifier by a significant amount.

Going back to the ROC curve of our classifier, a simple calculation would show that it is not realistic to expect a reasonable system level performance from a single-frame classification only. As mentioned previously, about 75 windows are inspected per frame, and given a processing rate of 10HZ we arrive at a number of 2.7 million classification queries in one hour of driving. Allowing for one false detection per 3 hours of driving, we would require a false alarm rate of 10^{-8} which is roughly 6 orders of magnitude better than what is displayed in the best ROC curve of Fig. 2. Such an improvement is not likely to happen by finding a better classifier or a better scheme for representing descriptors (local or global). The key therefore lies in the integration of additional cues measured over time (dynamic gait, motion parallax, stability of re-detection measures), situation specific features (such as leg positions at certain poses), and most importantly via building up additional object categories consisting of vehicles (both in motion and stationary) and stationary background structure such as poles, trees, guardrails, lane markings and so forth.

The details of the system level integration and the extraction of the additional cues are beyond the scope and space limitations of this paper. However, we will briefly present below some of the performance results of the

complete system.

4.1 The Complete System Performance

As mentioned in Section 2, the inspected windows which pass the single-frame classification stage undergo a multi-frame approval process. The accuracy requirements of the final decision depends on the location of the pedestrian and whether the pedestrian is stationary or moving laterally (inwards towards the host vehicle path). The most stringent requirements are set on inward moving pedestrians and on in-path stationary pedestrians. For these situations the false alarm rate should be less than 1 per 3 hours of driving with a detection rate of above 95%. Accuracy requirements for out-of-path stationary pedestrians are set for a false alarm rate of 360 false positives per hour of driving (roughly 1 per 10 seconds of driving) at a detection rate of 95%. As mentioned earlier, out-of-path stationary pedestrians are detected for purpose of advanced lock-in in order to minimize the detection delays in case the pedestrian decides to move in an inward motion — no external action is expected from the system upon detection of stationary out-of-path pedestrians.

We collected performance statistics over 5 hours of daytime driving in dense city traffic (mostly in downtown Tokyo and Jerusalem) under bright illumination with normal weather conditions. Although weather conditions were normal, bright illumination (as opposed to over-cast sky) introduces much difficulty to the detection process as shadows and highlights are emphasized in the scene and create in turn unstable contrast changes over the image and make the exposure control quite challenging due to the high dynamic range of the scene. The detections were divided into the following categories: (i) inward moving pedestrians, (ii) stationary pedestrians in-path, and (iii) stationary pedestrians out-of-path. Pedestrians moving longitudinally were counted as stationary. Stationary pedestrians out-of-path which were occluded (such as by parking vehicles and other obstructions) were not considered. The detection rate of inward-moving pedestrian stands on 96% with 1 false positive created during a host vehicle turning maneuver. The average delay for inward moving pedestrians at the range up to 15m was 4.6 frames (the minimal delay stands at 4 frmaes), 11.2 frames for 15m–25m and 21.7 frames for pedestrians at the range of 25m–30m. The detection rate of stationary in-path was calculated from a 1 hour drive in a busy district of Tokyo (see Fig. 4a,b) stands on 93% with 3 false positives. The detection rate of out-of-path pedestrians was determined from the same 1 hour session and stands on 85% with 102 false positives.

5 Summary

We have presented the functional requirements and architecture of a pedestrian detection system targeting on-board driving assistance applications. We presented our approach for the single-frame classification stage which is based on a novel scheme of breaking down the class variability by repeatedly training a set of relatively simple classifiers on clusters of the training set. Together with a shift-invariant local description of image sub-regions and discriminant integration using Adaboost we have obtained a powerful classifier that outperforms the leading approaches (for which a detailed description exists and can be re-produced). One of the key points made in this work is the observation that it is not realistic to expect a reasonable system level performance using single-frame classification only. The path from single-frame to system level performance must include the integration of additional cues measured over time (dynamic gait, motion parallax, stability of re-detection measures), situation specific features (such as leg positions at certain poses), and most importantly via building up additional object categories consisting of vehicles (both in motion and stationary) and stationary background structure such as poles, trees, guardrails, lane markings and so forth. The experimental results of our system so far indicate that for some of the functions (such as inward moving pedestrian detection) the performance is satisfactory for daytime and normal weather conditions, and for the remaining functionalities the gap which remains is relatively small for meeting a daytime normal weather specification.

References

[1] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifier. In *Proc. 5th Workshop on Computational Learning Theory*, pages 144–152, 1992.

[2] A. Broggi, M. Bertozzi, A. Fascioli and M. Sechi. Shape Based Pedestrian Detection. In *IEEE Inteligent Vehicle Symposium (IV2000)*, pp. 215-220, Dearborn, 2000.

[3] R. Cutler and L. Davis "Robust Real-Time Periodic Motion Detection, Analysis and Application" *IEEE Trans Patt. An. Mach. Int.*,, Vol 22(8), pp. 781-796, 2000.

[4] A. Efros, A. Berg, G. Mori and J. Malik "Recognizing Action at a Distance" *IEEE International Conference on Computer Vision (ICCV)*, pp. 726-733, 2003.

[5] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In Proceedings of International Conference on Machine Learning (ICML), pp. 148–156, 1996.

[6] D. Gavrila "Pedestrian Detection from a Moving Vehicle" *Proc. of the European Conference on Computer Vision (ECCV)*, pp. 37-49, Dublin, 2000.

[7] A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(3):55-67, 1970.

[8] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. In IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 23:349-361, April 2001.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004.

[10] O. Mano, G. Stein, E. Dagan and A. Shashua. Forward Collision Warning with a Single Camera In *IEEE Intelligent Vehicles Symposium (IV2004)*, June. 2004, Parma Italy.

[11] R. Okada, Y. Taniguchi, K. Furukawa and K. Onoguchi. Obstacle detection using projective invariant and vanishing lines. In *International Conference on Computer Vision* (ICCV), 2003.

[12] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna and T. Poggio Pedestrian detection using wavelet templates. In *Computer Vision and Pattern Recognition* (CVPR), June 1997.

[13] A. Shashua and N. Navab. Relative affine structure: Canonical model for 3D from 2D geometry and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (PAMI), 18(9):873–883, 1996.

[14] G. P. Stein, O. Mano and A. Shashua. A Robust Method for Computing Vehicle Ego-motion In *IEEE Intelligent Vehicles Symposium (IV2000)*, Oct. 2000, Dearborn, MI.

[15] G. P. Stein, O. Mano and A. Shashua, "Vision-based ACC with a Single Camera: Bounds on Range and Range Rate Accuracy" *IEEE Intelligent Vehicles Symposium (IV2003)*, June 2003, Columbus, OH.

[16] P. Reisman, O. Mano, S. Avidan and A. Shashua. Crowd Detection in Video Sequences In *IEEE Intelligent Vehicles Symposium (IV2004)*, June. 2004, Parma Italy.

[17] P. Viola, M. Jones and D. Snow "Detecting Pedestrians using Patterns of Motion Appearance" *IEEE International Conference on Computer Vision (ICCV)*, pp. 734-741, 2003.